

Climate Modeling at the Petaflop Scale Using Semi-custom Computing

Lenny Oliker, John Shalf, Michael Wehner

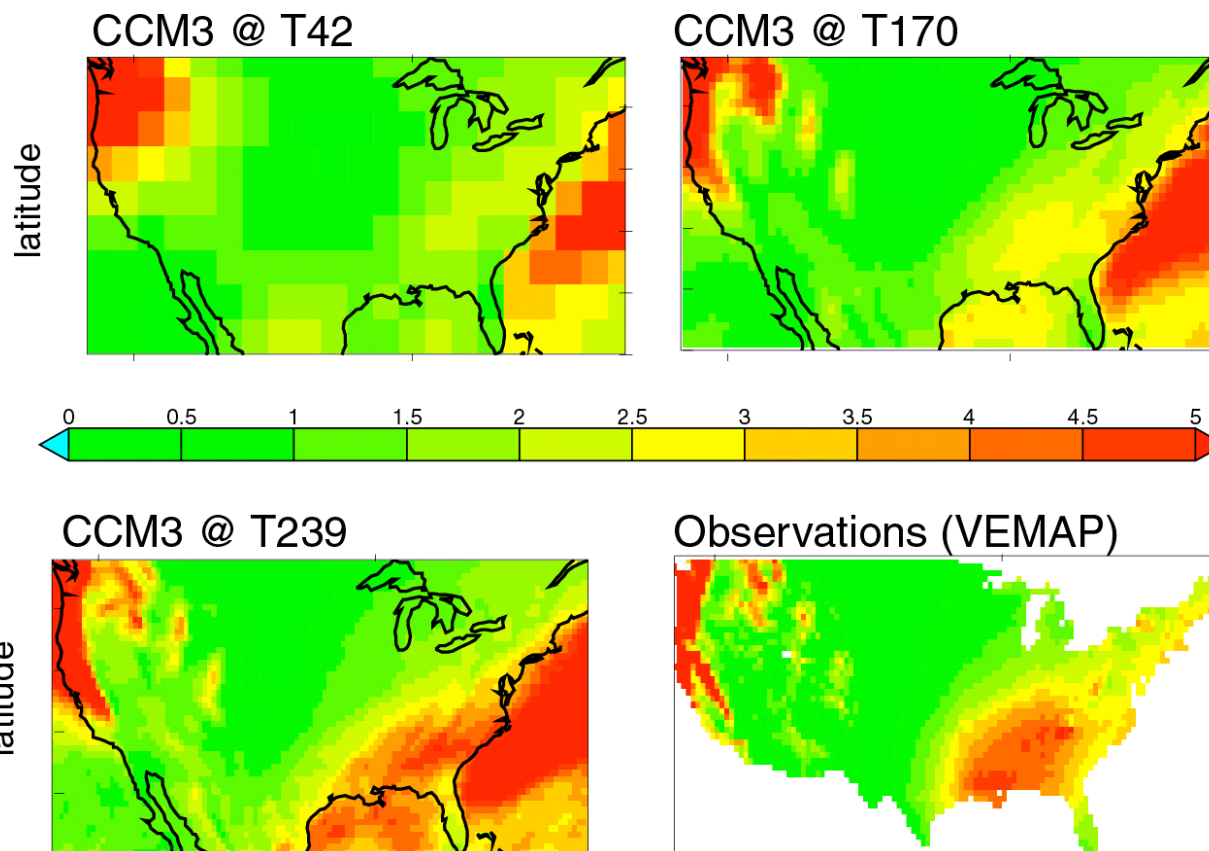
**Computational Research Division
National Energy Research Scientific Computing Center
Lawrence Berkeley National Laboratory
{loliker,jshalf,mwehner}@lbl.gov**

Motivations

- ❖ Accurately modeling climate change is one of the most critical challenges facing computational scientists today
 - Study anthropogenic climate change
 - Ramifications in trillions of dollars
- ❖ Current horizontal resolutions fail to resolve critical phenomena important to understanding the climate systems
 - Topographic effects: Both local and large scale
 - Tropical cyclones
 - At km-scale, important processes currently parameterized will be resolved
- ❖ We conduct speculative exploration of the computational requirements at ultra-high resolutions
 - Consider current technological trends
 - Explore alternative approaches to design semi-custom HPC solution
 - Show such calculations are reasonable within a few years time
 - Provide guidance to design of hardware/software to achieve goal
- ❖ Km-scale model would require significant algorithmic work as well as unprecedented levels of concurrency

Effects of Finer Resolutions

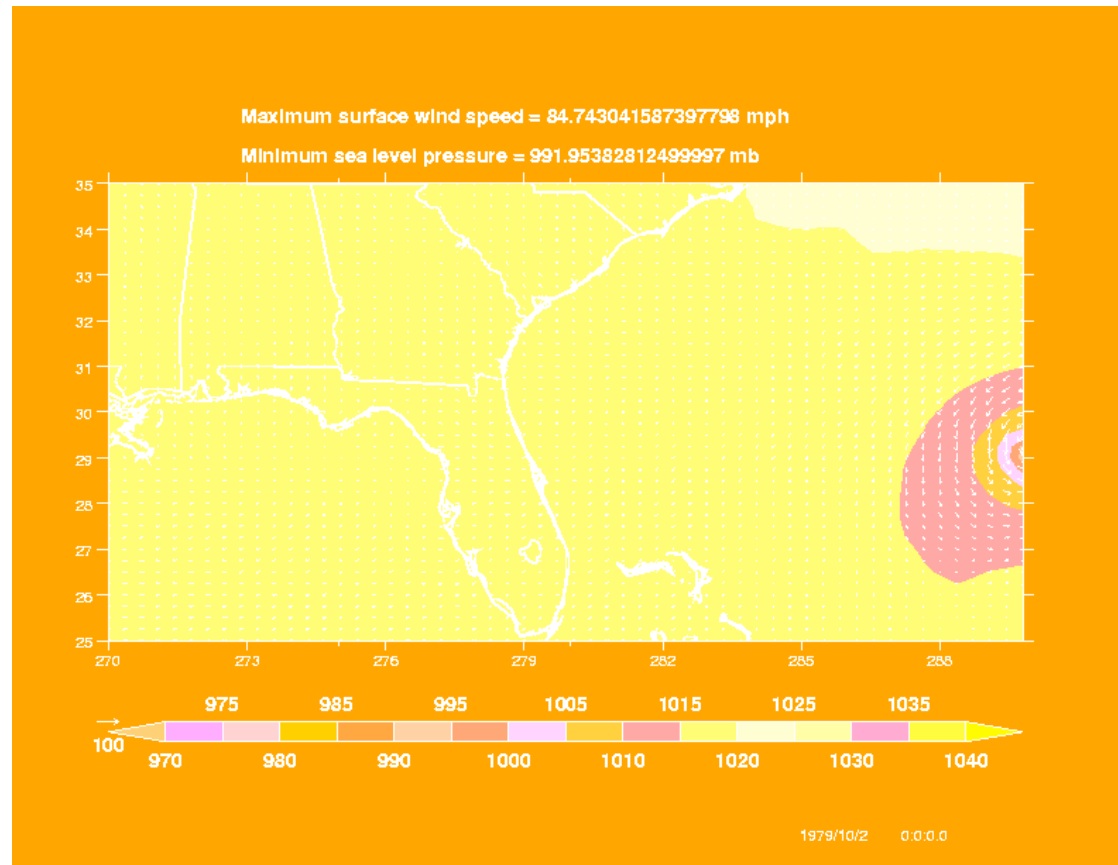
DJF Precipitation



Duffy, et al

Enhanced resolution of mountains yield
model improvements at larger scales

Pushing Current Model to High Resolution



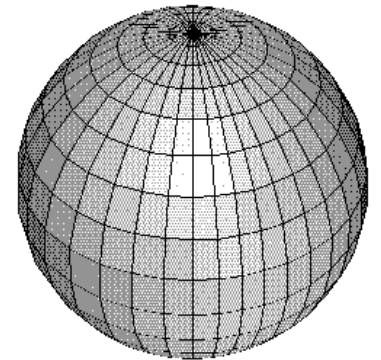
20 km resolution produces reasonable tropical cyclones

Kilometer-scale fidelity

- ❖ Current cloud parameterizations break down somewhere around 10km
 - Deep convective processes responsible for moisture transport from near surface to higher altitudes are inadequately represented at current resolutions
 - Assumptions regarding the distribution of cloud types become invalid in the Arakawa-Schubert scheme
 - Uncertainty in short and long term forecasts can be traced to these inaccuracies
- ❖ However, at ~2 or 3km, a radical reformulation of atmospheric general circulation models is possible:
 - Cloud system resolving models replace cumulus convection and large scale precipitation parameterizations.
 - Will this lead to better global cloud distributions

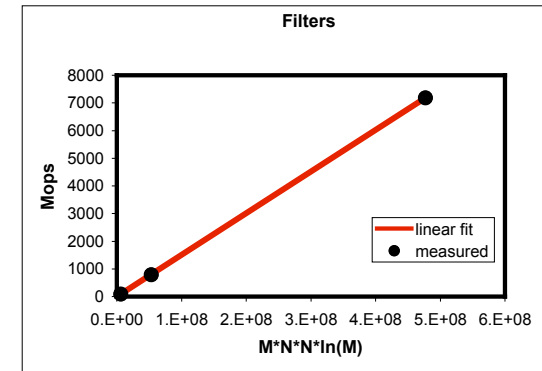
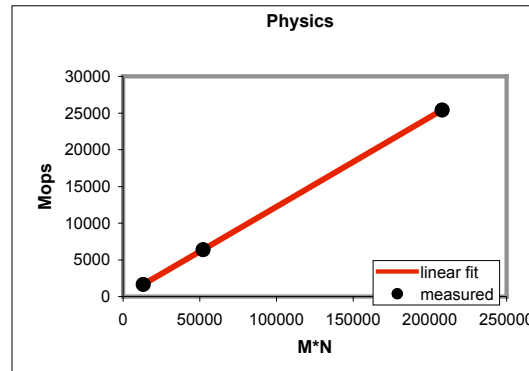
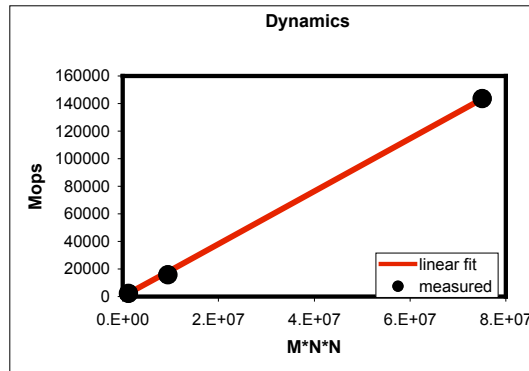
Extrapolating fvCAM to km Scale

- ❖ fvCAM: NCAR Community Atmospheric Model version 3.1
 - Finite Volume hydrostatic dynamics (Lin-Rood)
 - Parameterized physics is the same as the spectral version
 - Atmospheric component of fully coupled climate model, CCSM3.0
- ❖ We use fvCAM as a tool to estimate future computational requirements.
- ❖ Exploit three existing horizontal resolutions to establish the scaling behavior of the number of operations per fixed simulation period.
- ❖ Existing resolutions (26 vertical levels)
 - “B” $2^\circ \times 2.5^\circ$ (200 km), “C” $1^\circ \times 1.25^\circ$ (100 km), “D” $0.5^\circ \times 0.625^\circ$ (50 km)
- ❖ Define: m = # of longitudes, n = # of latitudes
- ❖ **Dynamics** - solves atmospheric motion, N.S. eqn fluid dynamics
 - Ops = $O(mn^2)$ Time step determined by the Courant (CFL) condition
 - Time step depends horizontal resolution (n)
- ❖ **Physics** - Parameterized external processes relevant to state of atmosphere
 - Ops = $O(mn)$, Time step can remain constant $\Delta t = 30$ minutes
 - Not subject to CFL condition
- ❖ **Filtering**
 - Ops = $O(m \log(m)n^2)$, addresses high aspect cells at poles via FFT
 - Allows violation of overly restrictive Courant condition near poles

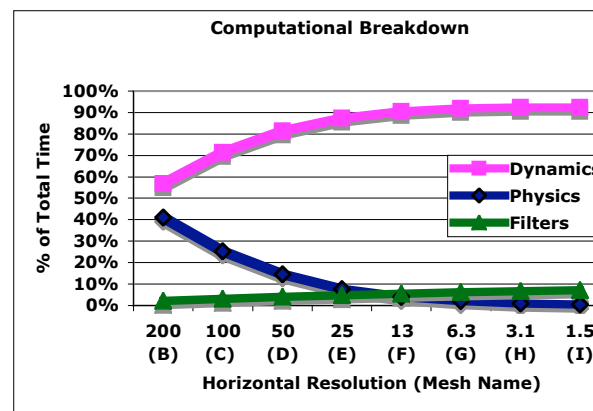


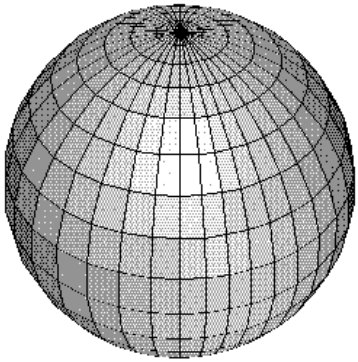
Extrapolation to km-Scale

Theoretical scaling behavior matches experimental measurements

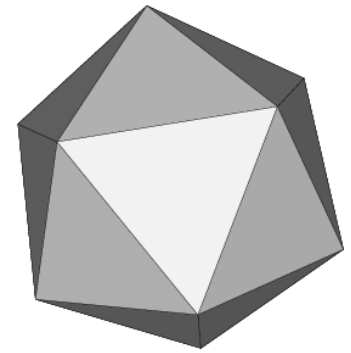


By extrapolating out to 1.5km, we see the dynamics dominates calculation time while Physics and Filters overheads become negligible

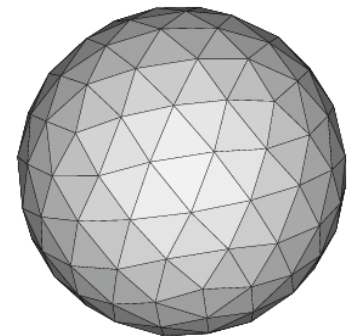
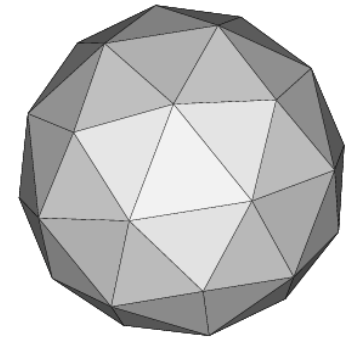




Caveats and Decomposition

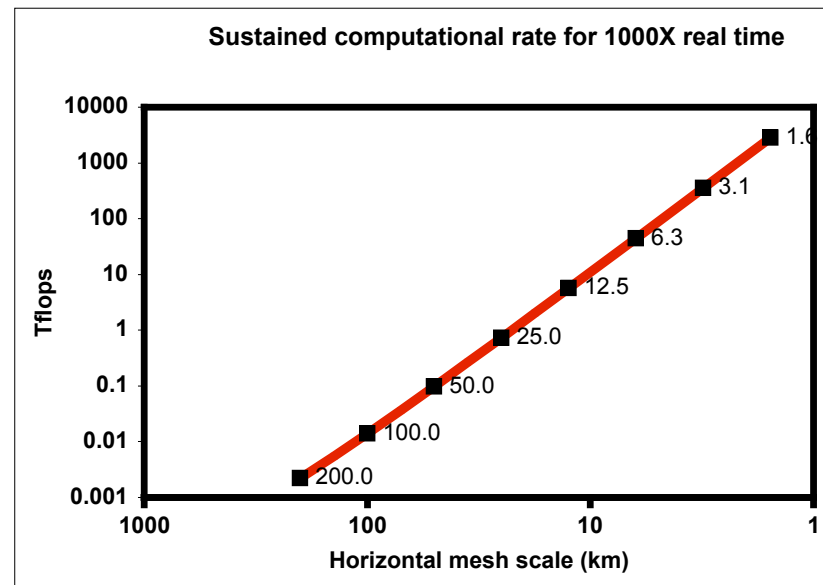


- ❖ Latitude-longitude based algorithm would not scale to 1km
 - Filtering cost would be only 7% of calculation
 - However the semi-Lagrangian advection algorithm breaks down
 - Grid cell aspect ratio at the pole is 10000!
 - Advection time step is problematic at this scale
- ❖ We thus make following assumptions:
 - Use Cubed sphere or icosahedral schemes for km-scale
 - Allows 2D decomposition as opposed to current 1D scheme
 - Computational costs at current resolutions are similar
 - Scaling behavior of dynamics is same as lat/long algorithms
 - Two horizontal spatial dimensions + Courant Condition (n^3)
- ❖ Physics time step can't stay constant if the subgrid scale parameterizations change.
 - Current cloud system resolving models use 10 second timestep.
 - Courant condition demands a 3.5 second timestep at km horizontal resolution for dynamics.
- ❖ Dynamics dominates the calculation

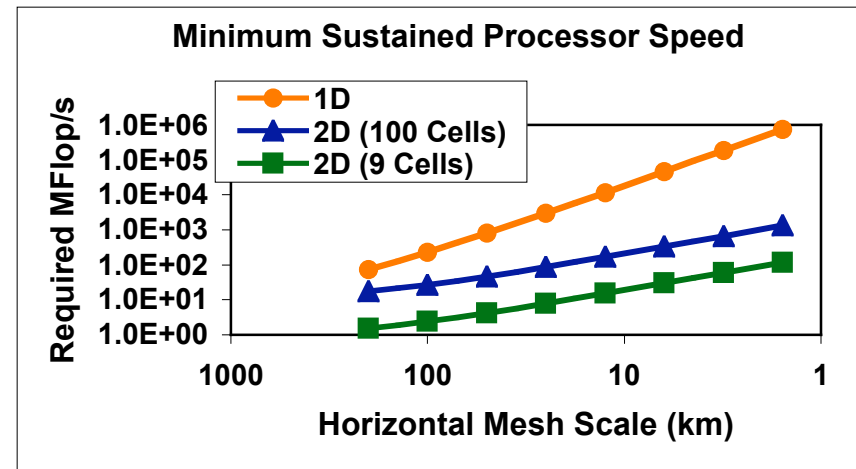
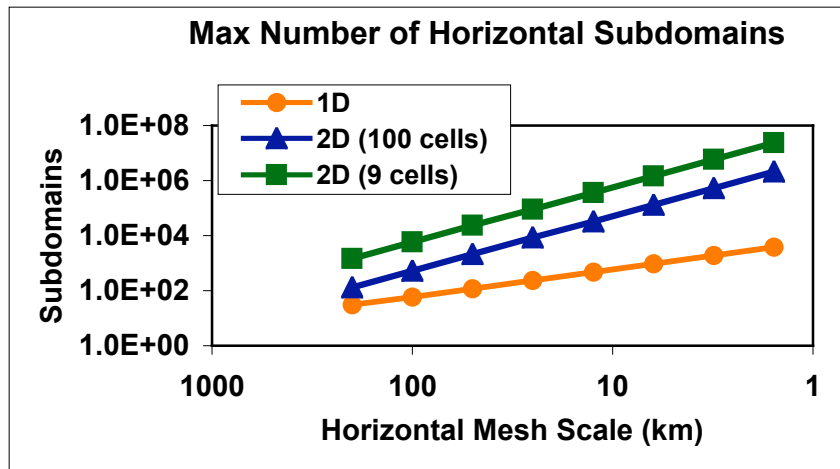


Sustained computational requirements

- ❖ A reasonable metric in climate modeling is that the model must run 1000 times faster than real time.
 - Millennium scale control runs complete in a year
 - Century scale transient runs complete in a month
 - For the moment hold the vertical layers constant @ 26
 - Weather prediction requires 10x realtime speedup
- ❖ At km-scale minimum *sustained* computational rate is 2.8 Petaflop/s
 - Number vertical layers will likely increase to 100 (4x increase) resulting in 10 Petaflop/s sustained requirement

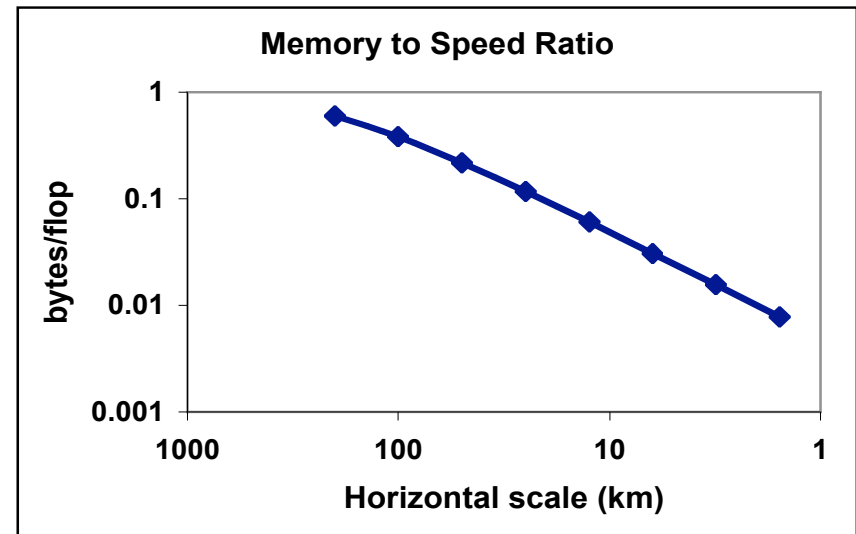
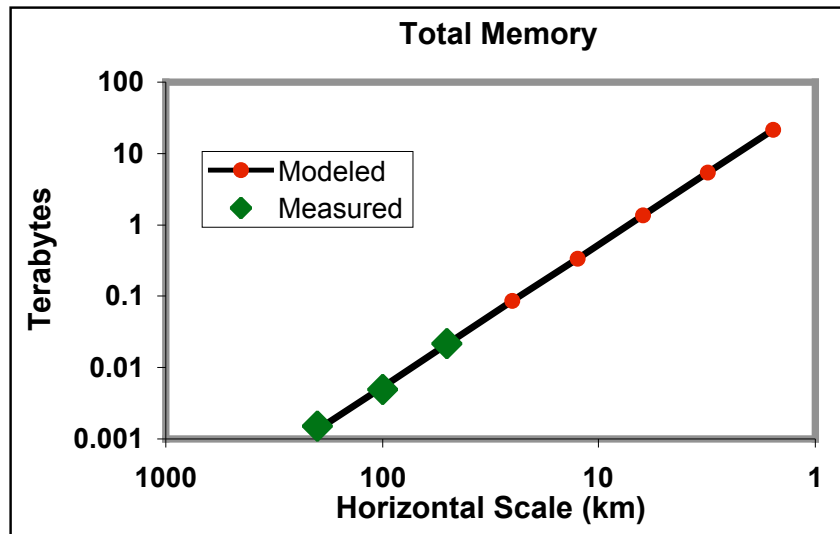


Processor scaling



- ❖ A practical constraint is that the number of subdomains is limited to be less than or equal to the number of horizontal cells
 - Using the current 1D approach is limited to only 4000 subdomains at 1km
 - Would require 1Teraflop/subdomain using this approach!
 - Number of 2D subdomains estimated using 3x3 or 10x10 cells
 - Can utilize millions of subdomains
 - Assuming 10x10x10 cells (given 100 vertical layers) = 20M subdomains
 - 0.5Gflop/processor would achieve 1000x speedup over realtime
 - Vertical solution requires high communication (aided with multi-core/SMP)
 - This is a lower bound in the absence of communication costs and load imbalance

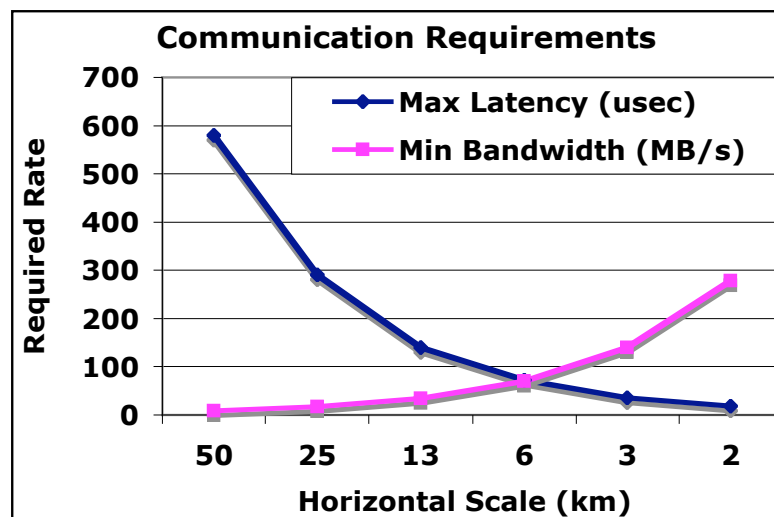
Memory Scaling Behavior



- ❖ Memory estimate at km-scale is about 25 TB total)
 - 100 TB total with 100 vertical levels
 - Total memory requirement independent of domain decomposition
- ❖ Due to Courant condition, operation count scales at greater rate than mesh cells - thus relatively low per processor memory requirement
 - Memory bytes per flop drop from 0.7 for 200km mesh to .009 for 1.5km mesh.
 - Using current 1D approach requires 6GB per processor
 - With 2D approach requires only 5MB per processor

Interconnect Requirements

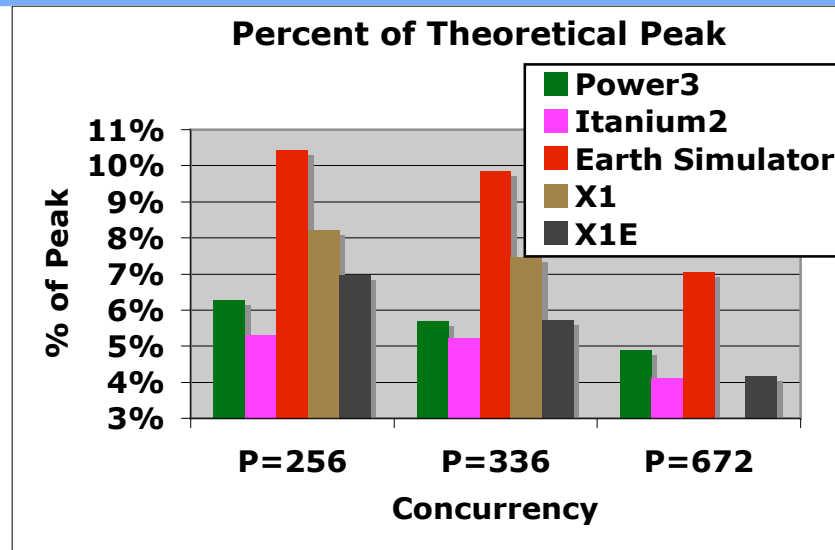
Data assumes 2D
10x10 decomposition
where only 10% of the
calculation is devoted
to communication



- ❖ Three factors cause sustained performance lower than peak:
 - Single processor performance, interprocessor communication, load balancing
- ❖ 2D case message size are independent on horizontal resolution, however in 1D case communication contains ghost cells over the entire range of longitudes
- ❖ Assuming (pessimistically) communication only occurs during 10% of calculation - not over the entire (100%) interval - increases bandwidth demands 10x
 - 2D 10x10 case requires: minimum 277 MB/s bandwidth and maximum 18 s latency
 - 1D case would require minimum of 256 GB/s bandwidth
- ❖ Note that the hardware/algorithm ability to overlap computation with communication would decrease interconnect requirements
- ❖ Load balance is important issue, but is not examined in our study

Today's Performance

Oliker, et al SC05



- ❖ Current state-of-the-art systems attain around 5% of peak at the highest available concurrencies
 - Note current algorithm uses OpenMP when possible to increase parallelism
- ❖ Thus peak performance of system must be 10-20x of sustained requirement

Strawman 1km Climate Computer

“I” mesh at 1000X real time

- .015°X.02°X100L (1.5km)
- 10 Petaflops *sustained*
- 100-200 Petaflops peak
- 100 Terabytes total memory
- Only 5 MB memory per processor
- 5 GB/s local memory performance per domain (1 byte/flop)
- 2 million horizontal subdomains
- 10 vertical domains (assume fast vertical communication)
- 20 million processors at 500Mflops each sustained
- 200 MB/s in four nearest neighbor directions
- Tight coupling of communication in vertical dimension

We now compare available technology in current generation of HPC systems

Declining Single Processor Performance

❖ Moore's Law

- Silicon lithography will improve by 2x every 18 months
- Double the number of transistors per chip every 18mo.

❖ CMOS Power

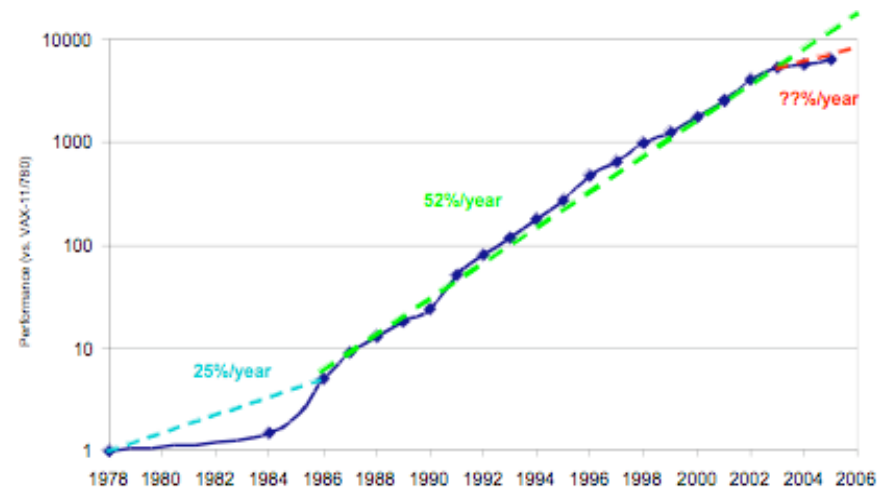
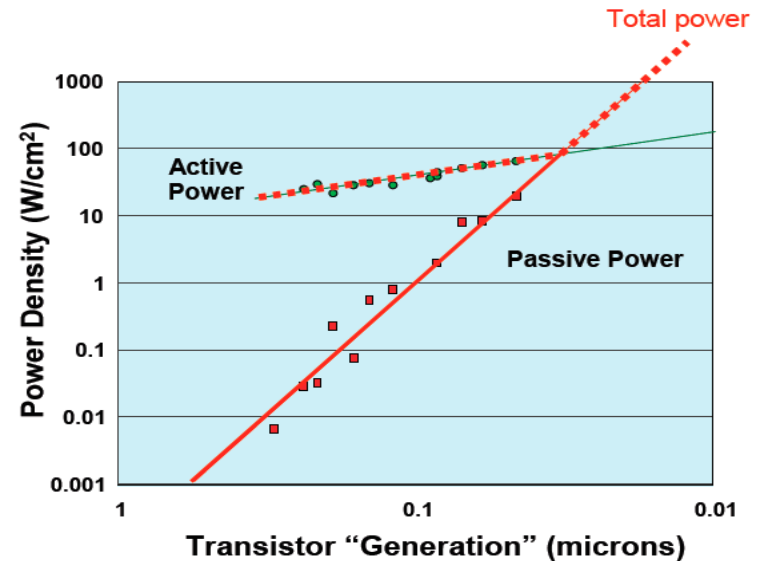
$$\text{Total Power} = \underbrace{V^2 * f * C}_{\text{active power}} + \underbrace{V * I_{\text{leakage}}}_{\text{passive power}}$$

- As we reduce feature size Capacitance (C) decreases proportionally to transistor size
- Enables increase of clock frequency (f) proportionally to Moore's law lithography improvements, with same power use
- This is called "Fixed Voltage Clock Frequency Scaling" (Borkar '99)

❖ Since ~90nm

- $V^2 * f * C \sim V * I_{\text{leakage}}$
- Can no longer take advantage of frequency scaling because passive power ($V * I_{\text{leakage}}$) dominates
- Result is recent clock-frequency stall reflected in Patterson Graph at right

❖ Multicore is here



SPEC_Int benchmark performance since 1978 from Patterson & Hennessy Vol 4.

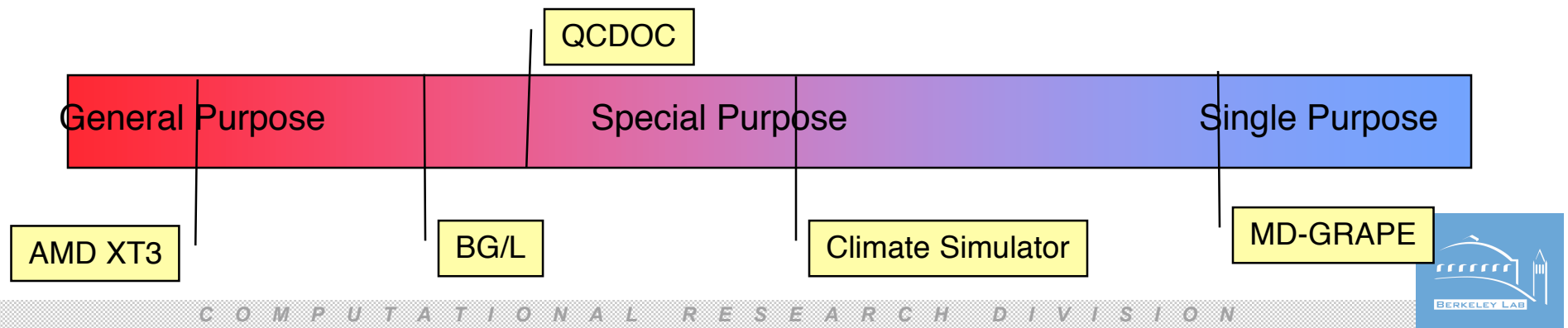


Learning from Embedded Market

- ❖ Desktop CPU market motivated to provide max performance at any cost.
 - Maximizing clock frequency
 - Long pipelines, complex o-o-o execution = extra power
 - Add features to cover virtually every conceivable application
 - Power consumption limited only by ability to dissipate heat
 - Cost around \$1K for high-end chips
- ❖ Embedded market motivated to maximize performance at min cost and power
 - Want cell phones that last forever on tiny battery and cost ~\$0
 - Specialized: remove unused features
 - *Effective* performance per watt is critical metric
- ❖ The world has changed
 - Clock frequency scaling has ended
 - At limited for cost effective air-cooled systems
 - Price point for desktops/portables dropping (portables dominate market)
 - For HPC, cost of power is exceeding procurement costs!
 - Technology from embedded market is now trickling up into server designs
 - Rather than traditional trickle down flow of innovations
- ❖ What will HPC learn from the embedded market?
 - Simpler, smaller cores
 - Many cores on chip (100's of cores, not 2,4,8)
 - Lower clock rates
 - More specialization to applications

Architectural Study of Climate Simulator

- ❖ We design system around the requirements of the km-scale climate code.
- ❖ Examined 3 different approaches
 - AMD Opteron: Commodity Approach - Lower efficiency for scientific applications offset by cost efficiencies of mass market
 - Popular building block for HPC, from commodity to tightly-coupled XT3.
 - Our AMD pricing is based on servers only without interconnect
 - BlueGene/L: Use generic embedded processor core and customize System on Chip (SoC) services around it to improve power efficiency for scientific applications
 - Power efficient approach, with high concurrency implementation
 - BG/L SOC includes logic for interconnect network
 - Tensilica: In addition to customizing the SOC, also customizes the CPU core for further power efficiency benefits but maintains programmability
 - Design includes custom chip, fabrication, raw hardware, and interconnect
- ❖ Continuum of architectural approaches to power-efficient scientific computing



Petascale Architectural Exploration

Processor	Clock	Peak/ Core (Gflops)	Cores/ Socket	Mem/ BW (GB/s)	Network BW (GB/s)	Sockets	Power <i>(based on current generation technology)</i>	Cost <i>(based on current market price)</i>
AMD Opteron	2.8GHz	5.6	2	6.4	4.5	890K	179 MW	\$1.8B
IBM BG/L	700MHz	2.8	2	5.5	2.2	1.8M	27 MW	\$2.6B
Climate computer	650MHz	2.7	32	51.2	34.5	120K	3 MW	\$75M

- ❖ AMD and BG/L based on list price
 - Of course discount pricing would apply, but extrapolation gives us baseline.
- ❖ Is it crazy to create a custom core design for scientific applications?
 - Yes, if the target is a small system.
 - In \$100M Petaflops system development costs are small compared to component costs.
 - In this regime, customization can be more power and cost effective than conventional systems.
 - Berkeley RAMP technology can be used to assess the design's effectiveness before it is built.
- ❖ Software challenges (at all levels) are a tremendous obstacle for any of these approaches.
 - Unprecedented levels of concurrency are required.
- ❖ This only gets us to 10 Petaflops *peak* - thus cost and power are likely to be 10x-20x more.
 - However, in ~5 years we can expect 8-16x improvement in power- and cost-efficiency.

Architectural Exploration using RAMP

What is Berkeley RAMP: Research Accelerator for Multiple Processors

- ❖ Sea of FPGAs linked together via hypertransport
- ❖ Provides enough programmable *gates* to simulate large chip designs
- ❖ Building community of “open source” hardware components (GateWare)
 - PPC4xx cores, Sun Niagara-1 netlists, Tensilica netlists
- ❖ Assemble gateway components (CPU and interconnects) using RDL (RAMP Description Language)
- ❖ Enables emulation of large clusters (100's or 1000's of nodes) using \$20K FPGA board.
 - Boots Linux - it looks like the *real* hardware to the software
 - Runs 100x slower than realtime, compared w/ million time slowdown of simulators
 - Can change HW parameters and explore new design on daily basis

We can explore climate supercomputer with RAMP

- ❖ Use Tensilica tools to generate netlists for Tensilica core design
 - Netlists describe list of logic gates and connections between them
 - Netlists is mapped and routed onto FPGAs to create working circuit
 - Protects CPU vendors intellectual property
- ❖ Use RDL to emulate subset of supercomputer (multi-core multi-socket design)
- ❖ Tensilica Open64 compilers can build code for specialized instruction set
- ❖ Build/run pieces of climate code on emulated machined to assess design

Conclusions

- ❖ Km scale resolution is a critical step towards more accurate climate models
 - Enables transition to more accurate physics-based cloud-resolving model
 - Supports unprecedented fidelity and accuracy for AGCM
- ❖ We extrapolate km-scale requirements to support such models
 - Developed specific requirements for sustained CPU, memory and interconnects
 - Provides guidance hardware and software designers
- ❖ Results show that riding the conventional technology curve will not enable us to reach these goals in the near future
 - Requires a more aggressive, power-efficient approach
- ❖ We suggest alternative approach to HPC designs by customizing hardware around the application -- not the other way around
 - Power-efficiency gains can be realized through semi-custom processor design
 - Otherwise energy costs for ultra-scale systems are likely to create a hard ceiling
 - We can reach our targets using near-term technology (without exotic technology)
 - Exploring opportunities to evaluate prototypes on Berkeley RAMP
- ❖ While custom hardware may not be cost-effective for mid-range problems, this approach may prove essential for handful of key Peta-scale applications
 - Future work will examine Fusion and Astrophysics
- ❖ Hardware, software, and algorithms are all equally critical, however HPC technology will probably be ready in advance of credible km-scale climate model
 - We must develop the algorithmic and architectural solutions simultaneously

Acknowledgements

- Art Mirin (Lawrence Livermore Laboratory)
- David Parks (NEC)
- Chris Rowen (Tensilica)
- Yu-Heng Tseng (National Taiwan University)
- Pat Worley (Oakridge National Laboratory)